SeeThruFinger: See and Grasp Anything with a Soft Touch

Fang Wan and Chaoyang Song *

Abstract

We present SeeThruFinger, a soft robotic finger with an in-finger vision for multi-modal perception, including visual perception and tactile sensing, for geometrically adaptive and real-time reactive grasping. Multi-modal perception of intrinsic and extrinsic interactions is critical in building intelligent robots that learn. Instead of adding various sensors for different modalities, a preferred solution is to integrate them into one elegant and coherent design, which is a challenging task. This study leverages the Soft Polyhedral Network design as a robotic finger, capable of omni-directional adaptation with an unobstructed view of the finger's spatial deformation from the inside. By embedding a miniature camera underneath, we achieve the visual perception of the external environment by inpainting the finger mask using E2FGV, which can be used for object detection in the downstream tasks for grasping. After contacting the objects, we use real-time object segmentation algorithms, such as XMem, to track the soft finger's spatial deformations. We also learn a Supervised Variational Autoencoder to enable tactile sensing of 6D forces and torques for reactive grasp. As a result, we achieve multi-modal perception, including visual perception and tactile sensing, and soft, adaptive object grasping within a single vision-based soft finger design compatible with multi-fingered robotic grippers.

Keywords: In-finger Vision, Visual Perception, Tactile Learning, Segment Anything, Inpainting

1. Introduction

Visual perception plays a central role in robotics to transform the unstructured environment into a structured embodiment for reasoning and learning in an object-centric manner.[1] While the classical method looks into the visual features from imaging to extract reusable information analytically,[2] recent development exploits data-driven methods by systematically integrating promptable segmentation tasks, pre-trained foundation models, and web-scale datasets to segment anything.[3] Such capability provides researchers with further insights to achieve object tracking,[4] object classification,[5] pose estimation,[6] occlusion inpainting,[7] and scene reconstruction[8] with pixel-level details in an automatic workflow, paving the path for a physical embodiment of machine intelligence through robot learning.[9] While it is preferable to introduce more modalities, vision-based perception remains a leading choice for many robot learning research and applications. However, due to the basic principles in optical physics, problems such as occlusion remain challenging in applications.

Vision-based sensory perception is probably the most widely adopted modality in modern robotics.[10–12] Besides bioinspiration from the human's hand-eye system,[13] robotic vision provides a unified representation of the 3D world into matrices of 2D pixels that have evolved rapidly in both theoretical foundations and engineering applications, which is further accelerated by recent advancements in data science and machine learning.[14] Occlusion is a practical problem caused by optical physics, such as line-of-sight, lighting, and reflection, which is inevitable for robot workstations that mainly adopted cameras as the only or primary sensor for extrinsic perception. To perform iterative visual verification during manipula-

*Fang Wan is with the School of Design, Southern University of Science and Technology, Shenzhen, China 518055. Chaoyang Song is the corresponding author with the Department of Mechanical and Energy Engineering, Southern University of Science and Technology, Shenzhen, China 518055. wanf@sustech.edu.cn, songcy@ieee.org

tion,[15] researchers usually need to add more cameras[16] or move the gripper to a particular pose to expose the fingers for grasp result checking, suffering trade-offs in hardware, integration, or time cost. While it is always possible to add more sensors or use more powerful ones, it remains a challenge to integrate multiple sensory modalities into a single unit for robotic manipulation.[17]

The robot manipulation problem is typically set up as a hand-eye system with a gripper ("hand") for handling physical contact with the objects and an external camera ("eye") for visual perception of the object-centric environment, with other components such as the manipulator, a desk-top, and target objects implied in a workstation.[18] Introducing tactile sensing on top of visual perception dramatically enhances flexibility in learning manipulation skills, formulating a coarse-to-fine hierarchy between globalized scene understanding and localized manipulation dynamics.[19] While researchers from other fields have developed various tactile sensors utilizing different sensing principles,[20] vision-based solutions show significant interest from robotic researchers in academia and applications,[21,22] probably due to a unified data structure, algorithmic processing, and electronic convenience from the vision sensors with compatible theories and principals in robotic vision and deep learning.

The visual occlusion problem is not unique to robots but also commonly faced by humans. Tactile sensory from the skin, especially those on the fingertips, is a biological answer to this problem with evolutionary evidence.[23] As the most favorable modality among roboticists, the research community is interested in developing vision-based tactile sensors to complement visual perception through contact.[24] However, due to the size of the camera board, the field of view, lenses, and wiring, it remains a design problem for vision-based tactile sensors to be seamlessly compatible with the various multi-fingered gripper designs. In-finger vision represents a general architecture of vision-based tactile sensing, usually designed with a piece of soft material handling physical interactions with the external environment through touch, then transforming the contact physics into material deformation, and finally captured by a miniature camera hidden underneath. In this way, unstructured contact physics is transformed into pixel-wise features in a dense resolution. Researchers have explored various design techniques, such as enclosed lighting in multiple colors,[25] artificial markers on the inside surface[26] or inside the soft material[27] and transparent material,[28] to enhance the imaging qualities before leveraging recent advancement of computer vision in algorithm, data, and computation.

[Figure 1 about here.]

There remains a research gap in design in the current literature to leverage the soft material adaptation in handling physical contact for enhanced manipulation while generating a rich image stream of the material deformations of the object-centric environment and contact physics simultaneously. We propose the SeeThruFinger architecture shown in Figure 1 to address this limitation, which features an in-finger vision for multi-modal perception in manipulation learning, capable of geometric adaptation, visual perception, and tactile sensing in a single design.

- **Soft Finger**: The soft finger adopts a variant of the Soft Polyhedral Network designs.[29] A miniature camera is embedded underneath to capture the whole-body deformation of the soft finger from the inside. One can easily redesign the adaptor to fix the soft finger to most industrial multi-fingered grippers with added mechanical benefits in geometric adaptation, size of grasping, and impact absorption at a low cost.

- **Scene Inpainting**: Before contact occurs, assuming no collision with any object, the finger covers a fixed area of pixels as a mask on the image, occluding the scene. We can use threshold segmentation against a simple background. We verified the effectiveness of using inpainted images for downstream tasks such as object detection and manipulation planning.

- **Tactile Sensing**: As contact begins, we can use the same finger mask obtained earlier and implement algorithms for real-time object segmentation, such as XMem,[4] to track the soft finger's spatial deformation and use it to train a Supervised Autoencoder against true labels obtained from an ATI nano25. Finally, we achieved real-time tactile sensing that can be used for reactive grasping to adjust the gripper pose for enhanced grasping.

2. Problem formulation

Motivated by the fact that the in-finger camera sees both the deformation of the soft network structure and the world outside, we learn a fusion of visual perception and tactile sensing via a single in-finger vision, making every pixel it sees contribute to one of the two perception modalities. We have a mask template $\mathbf{M_0}$ depicting the area blocked by the finger's network without external load. When the soft finger interacts with objects in robotic tasks such as bin picking, the in-finger camera captures an image $\mathbf{I_t} \in \mathbb{R}^{360\times 640}$ at each timestep t. We track the time evolution of the mask $\mathbf{M_t}$. The goal is to learn a visual perception model $f(\{\mathbf{I_t} - \mathbf{M_t}\}_{t\in[t_1,t_2]})$ to understand the scene of workspace and a tactile perception model $\mathbf{FT_t} = g(\mathbf{M_t}, \mathbf{M_0})$ to infer instantaneous force and torque exerted on the center of soft finger base, where $\mathbf{FT_t} = [f_x, f_y, f_z, \tau_x, \tau_y, \tau_z]_t$. This implies the 2D mask contains adequate information on the 3D soft deformation, which further determines the resultant force and torque at the base of the soft finger.

3. Methods

3.1. Design Integration of the SeeThruFinger

This study adopts the DeepClaw workstation[18] built with aluminum extrusion, housing a UR10e from Universal Robots on a pedestal and a tabletop in front covered by fabrics in black color. As shown in Figure 2(a), although a camera is fixed on the post, it is unplugged as we will implement visual perception for the scene through inpainting using in-finger vision from the SeeThruFinger. Figure 2(b) shows a detailed view of the two-fingered gripper (Model AG-160-95 by DH-Robotics) with its fingertips replaced by the SeeThruFingers. Each SeeThruFinger contains a soft, omni-adaptive finger designed based on a variant of the Soft Polyhedral Networks,[29] mounted on a camera housing 3D-printed by UV-curable transparent resin (Somos WaterShed XC 1112), where a miniature camera (Chengyue WX605 from Weixinshijie) is fixed inside, as shown in Figure 2(c). One can customize the base plate to install the SeeThruFinger on FT sensors for testing or on a gripper as fingertips for soft and adaptive grasping. Compared to our previous work,[29] the SeeThruFinge features an enhanced design with a larger contact surface for grasping, enabled by the two vertices on top and no marker inside the finger. Figure 2(d) shows the objects used for testing in this study.

[Figure 2 about here.]

3.2. The SeeThruFinger Architecture with Robot Learning

This study proposes the SeeThruFinger architecture for learning visual perception and tactile sensing simultaneously via the in-finger vision with a soft touch, as shown in Figure 1. Besides omni-directional adaptation for grasping, the SeeThruFinger also features an in-finger vision with direct line-of-sight to the external environment but suffers from occlusions by the finger network. By segmenting the finger mask either through simple thresholding or with clickable interaction with SAM,[3] we implement the inpainting algorithm, i.e., E2FGV,[30] to generate the occluded scene and objects for visual perception. Then, we performed object detection using the inpainted image, i.e., Real-Time Detection Transformer (RT-DETR),[31] generating object class and its planar coordinates (x, y) on the tabletop for manipulation planning. During its first attempt, the gripper will attempt to pick up the object at a fixed pose and depth z_0 relative to the robot. While closing the fingers, the same SeeThruFinger detects if it has contact with the object by inferring tactile feedback in 6D forces and torques based on the finger's whole-body deformation. Usually, contact normal to the finger surface will cause the finger to deform adaptively for enhanced grasping results.[32] Due to the network design, the finger can deform twistedly to generate torque about the z axis. When τ_z is detected to be larger than 0.05 Nm, the object orientation is severely different from the gripper's initial pose, suggesting a less ideal grasping that can be optimized through regrasping by turning the wrist joint reversely to reduce τ_z for an enhanced adaptation and grasping robustness. We implemented a Supervised Variational Autoencoder (SVAE) to enable tactile sensing using SeeThruFinger's design by tracking the finger's whole-body deformation after contact.

3.3. *Supverised Variational Autoencoder for Tactile Sensing*

Previous work shows that it is possible to use Convolutional Neural Networks for tracking soft finger interactions,[29] which suffer from a less accurate prediction in 6D forces and torques. Furthermore, it requires a uniform background or enclosed lighting, a similar problem shared by many other vision-based tactile sensors.[26] Adding an ArUco marker inside the finger is a working solution but would introduce extra rigid plates and designs that inevitably limit its soft adaptation capabilities. In this study, we leverage the latest development in video object segmentation, i.e., XMem,[4] to track the soft finger's spatial deformations in real time, even if the background is noisy. As shown in Figure 1, using the tracked deformation as the input, we collected true labels of 6D forces and torques at the finger base using nano25 by ATI and trained an SVAE model to learn the latent representation of the soft finger's geometric adaptation as well as the 6D forces and torques simultaneously. As shown in Figure 2(e), the real-time mask of the soft finger $\mathrm{M_t}$ and the initial template $\mathrm{M_0}$ are stacked and go through the ResNet18 encoder and decoder module to reconstruct the original image. At the same time, we design an auxiliary supervised task of force and torque regression from the learned latent geometry representation. The overall loss consists of reconstruction loss, the MSE loss of force/torque prediction, and the Kulback-Leibler divergence. All codes are available at https://github.com/ancorasir/SeeThruFinger.

4. Experiments and Results

The following experiments demonstrate our proposed approach combining visual detection and tactile control in a table cleaning task. As shown in Figure 3(a), although both SeeThruFingers on the gripper are functional, we only used the one whose viewing range partially covers the table at every frame throughout the following experiments.

[Figure 3 about here.]

4.1. *In-Finger Scene Inpainting and Object Detection with Static Occlusion*

As shown in Figure 3(b), we designed a head-up movement of the gripper for the in-finger camera to capture a short video of about 1 second in which its viewing range sweeps across the tabletop. The first row of Figure 3(c) shows the in-finger vision of the scene without attaching the soft finger. The second row with the soft finger occluding the scene, and the third for the corresponding frames of the inpainted scene. Only the right half of the image with objects is inpainted for efficient processing. Notice that the mask stays the same, and the environment outside changes.

For inpainting, we first obtain the mask template $\mathrm{M_0}$ in Figure 3(d) using threshold segmentation against a clean background. Results show that the soft finger occludes about 53.6% of the image pixels. Alternatively, we can use SAM to obtain the mask by clicking for comparison in an actual scene. Figure 3(e) shows the inpainting operation implemented with E2FGV[30] using raw image stream from in-finger vision and the finger mask from threshold segmentation, generating the scene with objects on the tabletop. The inpainted images managed to reconstruct sufficient details of the occluded objects and the scene, as shown in the dashed white boxes in Figure 3(c). We tested the feasibility of the inpainted scene by performing object detection using RT-DETR[31] based on the inpainted scene, producing object classification (%) and bounding box center (x, y) for manipulation planning. Figure 3(f) is an example of the object detection performance, which is acceptable for simple grasping. Although it would be interesting to test with other more advanced pose estimation algorithms, we intentionally used a simpler algorithm that generates only the center of the bounding box to test reactive grasping with soft adaptation[32] and tactile sensing[33] using the SeeThruFinger.

The proposed SeeThruFinger effectively reduces a hand-eye system's hardware complexity using a single in-finger camera to achieve visual and tactile perception, which usually requires two separate sensors with added hardware cost and complexity in system integration in previous work. Compared to existing solutions of vision-based tactile sensing, our design does not require an enclosed chamber for in-finger perception. It retains a reasonable partial view of the scene, making it possible for scene inpainting. Even with GelSight type of vision-based tactile sensors where transparent elastomers are used, the view is still blocked by the external coating on the surface[26] or suffers from a blurry image of the scene that can only be detected at a very close range,[28] insufficient for visual perception of the whole table-top.

4.2. *In-Finger Tactile Sensing with SVAE by Tracking Soft Deformations using XMem*

The SeeThruFinger is capable of spatial adaptation against different object geometries (Figure 4(a)). As contact begins in Figure 4(b), the in-finger vision captures the finger's whole-body deformations for tactile sensing. However, when installed on the gripper, due to the changing background visible through gaps between the network structure, it is challenging to segment the soft network efficiently when deforming simultaneously. To address this challenge, as shown in Figure 4(c), we implemented real-time (30 Hz) object tracking of the soft finger using XMem[4] while deforming based on the same finger mask obtained earlier. We evaluated this mask segmentation over a test dataset of 1,000 samples with the ground truth masks. The region (J) and boundary (F) measures[34] are 0.975 and 0.997, respectively, showing excellent reliability in tracking the soft finger's spatial deformations.

[Figure 4 about here.]

Based on the segmented images of the soft finger during deformation, we implement a vision-based tactile learning algorithm using SVAE. We collected the training data using the setup in Figure 4(e). The in-finger camera and FT sensor were connected to a laptop for data streaming. We collected 40,000 synchronized pairs of image and FT readings by compressing the two soft fingers from various directions and contact locations. These two fingers are later integrated into a two-finger gripper and execute the table-cleaning task. We used an 8:2 train-valid split and trained the SVAE model for 50 epochs. To avoid excessive good validation scores due to the potential sample dependence in the train and validation dataset, we separately collected a test dataset of 1,000 samples.

The average time for a singer inference is 1.9 ms on an NVidia 3080 Ti Laptop GPU. The Mean Absolute Errors of the validation and test dataset are reported in Table 1, indicating good generalization of the learned model. Figure 4(f) compares the prediction and ground truth time in series in the test dataset. Figure 4(g) shows the scatter plot of such comparison in the validation dataset with excellent R^2 scores over 0.985 in all components except f_z. The relatively weak perforce in f_z is due to the limited adaptability along the z-axis. We also evaluate the resultant force in the xy plane and find the mean magnitude and orientation error is 0.286 N and 2.6 degrees, respectively.

[Table 1 about here.]

The integrated design is another benefit of the SeeThruFinger as a vision-based soft robotic finger instead of a standalone sensor. The proposed SeeThruFinger features a 3D metamaterial design as a soft networked structure,[29] capable of omni-directional adaption as an integrated finger with added benefits such as geometric adaptation. By changing the adaptors, one can directly enjoy the compounded benefits of soft adaptation, scene inpainting, and tactile sensing in another gripper.

4.3. *Reactive Grasp Learning using SeeThruFinger with In-Finger Vision*

Here, we combine the above results into a SeeThruFinger architecture. Figure 5(a) shows the recorded history of 6D FT sensing using the SeeThruFinger for reactive grasping, where the six peaks in f_x indicate grasping of the six test objects. The inpainted scene using SeeThruFinger is shown in Figure 5(b), where the inpainted area managed to reconstruct the objects' overall details. It should be noted that the reconstruction of the occluded geometry is well reconstructed. However, most surface textures are generated blurry but are still recognizable and sufficiently accurate to reflect their actual textures. Figure 5(c) shows object detection results using RT-DETR based on the inpainted scene, showing bounding boxes for each detected object.

[Figure 5 about here.]

The most interesting result is the regrasping of the Bosche electrical screwdriver shown in Figure 5(d). By inspecting the detailed FT history, we identified two regrasping attempts performed by the SeeThruFinger based on tactile sensing and adaptive grasping. As shown in Figure 5(e), when the SeeThruFinge performed the first grasping attempt in (i), the finger immediately detected the τ_z to be greater than 0.05 Nm in the negative direction. Then, (ii) the first reactive grasping begins rotating the gripper about the z axis to reduce τ_z. During the process, (iii) the screwdriver is turned to another side suddenly and temporarily loses contact with the finger, as both f_x and f_y dropped to 0 suddenly. Even so, (iv) the gripper keeps closing, attempting to detect contact with the screwdriver again to keep increasing τ_z for an enhanced grasping pose. Next, (v) the SeeTruFinger managed to detect contact with the screwdriver again, marking the starting point of the second reactive grasping attempt. Interestingly, SeeThruFinger repeated the same procedure by performing the second reactive grasping (vi) by rotating the gripper about the z axis until the detected τ_z is reduced towards 0. As a result, (vii) the screwdriver is secured between the fingers and moved toward the drop box. Finally, (viii) the SeeThruFinger dropped the screwdriver with zero values detected from tactile sensing using in-finger vision.

5. Conclusion and Limitations

This paper presents a design and learning approach for the SeeThruFinger architecture that simultaneously achieves soft and adaptive grasping, visual perception via scene inpainting, and reactive grasping via tactile sensing using a single in-finger vision sensor and a soft finger network design. Due to the unique design with a fully exposed visual perception of the contact interactions and the integration of learning algorithms, we achieved multi-modal perception, including globalized scene segmentation and object detection, and localized, real-time tactile sensing with high accuracy, while providing passive adaption in omni-directions during grasping with the soft network design. A direct result of the proposed method is the possibility of removing the external camera completely while providing physical compliance, visual perception, and tactile sensing at the time using just one in-finger vision, which is possible based on what we have reported in this study.

Our work has several limitations. Even though the XMem is reasonably fast (30 Hz), it is still far from the chosen camera's 330 Hz framerate. The whole pipeline is not as efficient yet, which requires further optimization for robotic interactions. In this work, we did not fine-tune the object detection model for the objects tested, and the object detection results show mislabeled object classes even though they can be detected. The learned tactile perception model is subject to a systematic shift when applied to a new finger not presented in the training dataset due to the slight difference in the initial template mask of each finger. Although we fed the initial template mask along with the instantaneous one, SVAE failed to encode the geometric deformation compared to the template mask, probably because it only saw two templates during training. Besides mitigating this systematic error by simple subtraction, a promising way is through data augmentation.

There are a few directions we would like to explore further in the future. One is underwater visual perception and tactile sensing using SeeThruFinger by translating models learned on land, aiming at dexterous and sensitive exploration underwater.[35] Due to the unique design, the SeeThruFinger can be directly used underwater if the camera housing is waterproofed, which is quite simple as there are no dynamic seals. In this study, only one of the SeeThruFingers' in-finger vision was used, but both are fully functional. It would be interesting to see if the stereo vision for 3D reconstruction or 6D pose estimation based on the inpainted scene from both in-finger visions would be possible, which is another exciting direction we would like to explore further.

Acknowledgements

This work was partly supported by the National Natural Science Foundation of China [Grant: 62206119], the Science, Technology, and Innovation Commission of Shenzhen Municipality [Grant: ZDSYS20220527171403009, JCYJ20220818100417038], the SUSTech-MIT Joint Centers for Mechanical Engineering Research and Education, Guangdong Provincial Key Laboratory of Human Augmentation and Rehabilitation Robotics in Universities.

Author Disclosure Statement

The authors declare that they have no competing interests.

References

[1] N. Sünderhauf, O. Brock, W. Scheirer, R. Hadsell, D. Fox, J. Leitner, B. Upcroft, P. Abbeel, W. Burgard, M. Milford, and P. Corke, "The Limits and Potentials of Deep Learning for Robotics," *The International Journal of Robotics Research*, vol. 37, no. 4-5, pp. 405–420, 2018.

[2] P. Meer, D. Mintz, A. Rosenfeld, and D. Y. Kim, "Robust Regression Methods for Computer Vision: A Review," *International Journal of Computer Vision*, vol. 6, pp. 59–70, 1991.

[3] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment Anything," 2023.

[4] H. K. Cheng and A. G. Schwing, "XMem: Long-Term Video Object Segmentation with an Atkinson-Shiffrin Memory Model," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*, pp. 640–658, Springer, 2022.

[5] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, 2017.

[6] E. Murphy-Chutorian and M. M. Trivedi, "Head Pose Estimation in Computer Vision: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 607–626, 2008.

[7] T. Yu, R. Feng, R. Feng, J. Liu, X. Jin, W. Zeng, and Z. Chen, "Inpaint Anything: Segment Anything Meets Image Inpainting," 2023.

[8] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[9] D. Howard, A. E. Eiben, D. F. Kennedy, J.-B. Mouret, P. Valencia, and D. Winkler, "Evolving Embodied Intelligence from Materials to Machines," *Nature Machine Intelligence*, vol. 1, no. 1, pp. 12–19, 2019.

[10] R. M. Murray, Z. Li, and S. Shankar, *A Mathematical Introduction to Robotic Manipulation*. CRC press, 1994.

[11] J. J. Craig, *Introduction to Robotics*. Pearson Educacion, 2006.

[12] K. M. Lynch and F. C. Park, *Modern Robotics*. Cambridge University Press, 2017.

[13] P. I. Corke and O. Khatib, *Robotics, Vision and Control: Fundamental Algorithms in MATLAB*, vol. 73. Springer, 2011.

[14] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[15] O. Kroemer, S. Niekum, and G. Konidaris, "A Review of Robot Learning for Manipulation: Challenges, Representations, and Algorithms," *Journal of Machine Learning Research*, vol. 22, no. 30, pp. 1–82, 2021.

[16] OpenAI, I. Akkaya, M. Andrychowicz, M. Chociej, M. Litwin, B. McGrew, A. Petron, A. Paino, M. Plappert, G. Powell, R. Ribas,

J. Schneider, N. A. Tezak, J. Tworek, P. Welinder, L. Weng, Q. Yuan, W. Zaremba, and L. M. Zhang, “Solving Rubik’s Cube with a Robot Hand,” *ArXiv*, 2019.

[17] S.-F. Zhang, J.-H. Zhai, B.-J. Xie, Y. Zhan, and X. Wang, “Multimodal Representation Learning: Advances, Trends and Challenges,” in *2019 International Conference on Machine Learning and Cybernetics (ICMLC)*, pp. 1–6, 2019.

[18] F. Wan, H. Wang, X. Liu, L. Yang, and C. Song, “DeepClaw: A Robotic Hardware Benchmarking Platform for Learning Object Manipulation,” in *2020 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, pp. 2011–2018, 2020.

[19] M. Lee, Y. Zhu, P. Zachares, M. Tan, K. Srinivasan, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, “Making Sense of Vision and Touch: Learning Multimodal Representations for Contact-Rich Tasks,” *IEEE Transactions on Robotics*, vol. PP, pp. 1–15, 03 2020.

[20] Y. Liu, R. Bao, J. Tao, J. Li, M. Dong, and C. Pan, “Recent Progress in Tactile Sensors and Their Applications in Intelligent Systems,” *Science Bulletin*, vol. 65, no. 1, pp. 70–88, 2020.

[21] Q. Li, O. Kroemer, Z. Su, F. F. Veiga, M. Kaboli, and H. J. Ritter, “A Review of Tactile Information: Perception and Action through Touch,” *IEEE Transactions on Robotics*, vol. 36, no. 6, pp. 1619–1634, 2020.

[22] S. Zhang, Z. Chen, Y. Gao, W. Wan, J. Shan, H. Xue, F. Sun, Y. Yang, and B. Fang, “Hardware Technology of Vision-Based Tactile Sensor: A Review,” *IEEE Sensors Journal*, vol. 22, no. 22, pp. 21410–21427, 2022.

[23] E. Macaluso and A. Maravita, “The Representation of Space Near the Body through Touch and Vision,” *Neuropsychologia*, vol. 48, no. 3, pp. 782–795, 2010.

[24] K. Shimonomura, “Tactile Image Sensors Employing Camera: A Review,” *Sensors*, vol. 19, p. 3933, 09 2019.

[25] H. Sun, K. J. Kuchenbecker, and G. Martius, “A soft thumb-sized vision-based sensor with accurate all-round force perception,” *Nature Machine Intelligence*, vol. 4, no. 2, pp. 135–145, 2022.

[26] W. Yuan, S. Dong, and E. H. Adelson, “Gelsight: High-resolution robot tactile sensors for estimating geometry and force,” *Sensors*, vol. 17, no. 12, p. 2762, 2017.

[27] C. Sferrazza and R. D’Andrea, “Design, motivation and evaluation of a full-resolution optical tactile sensor,” *Sensors*, vol. 19, no. 4, 2019.

[28] F. R. Hogan, J.-F. Tremblay, B. H. Baghi, M. Jenkin, K. Siddiqi, and G. Dudek, “Finger-STS: Combined Proximity and Tactile Sensing for Robotic Manipulation,” *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10865–10872, 2022.

[29] F. Wan, X. Liu, N. Guo, X. Han, F. Tian, and C. Song, “Visual Learning towards Soft Robot Force Control using a 3D Metamaterial with Differential Stiffness,” in *Proceedings of the 5th Conference on Robot Learning* (A. Faust, D. Hsu, and G. Neumann, eds.), vol. 164 of *Proceedings of Machine Learning Research*, pp. 1269–1278, PMLR, 08–11 Nov 2022.

[30] Z. Li, C.-Z. Lu, J. Qin, C.-L. Guo, and M.-M. Cheng, “Towards an End-to-End Framework for Flow-Guided Video Inpainting,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17541–17550, 2022.

[31] W. Lv, S. Xu, Y. Zhao, G. Wang, J. Wei, C. Cui, Y. Du, Q. Dang, and Y. Liu, “DETRs Beat YOLOs on Real-time Object Detection,” 2023.

[32] F. Wan, H. Wang, J. Wu, Y. Liu, S. Ge, and C. Song, “A Reconfigurable Design for Omni-Adaptive Grasp Learning,” *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4210–4217, 2020.

[33] L. Yang, X. Han, W. Guo, F. Wan, J. Pan, and C. Song, “Learning-Based Optoelectronically Innervated Tactile Finger for Rigid-Soft Interactive Grasping,” *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3817–3824, 2021.

[34] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 724–732, 2016.

[35] O. Khatib, X. Yeh, G. Brantner, B. Soe, B. Kim, S. Ganguly, H. Stuart, S. Wang, M. Cutkosky, A. Edsinger, P. Mullins, M. Barham, C. R. Voolstra, K. N. Salama, M. L'Hour, and V. Creuze, "Ocean One: A Robotic Avatar for Oceanic Discovery," *IEEE Robotics and Automation Magazine*, vol. 23, no. 4, pp. 20–29, 2016.

List of Figures

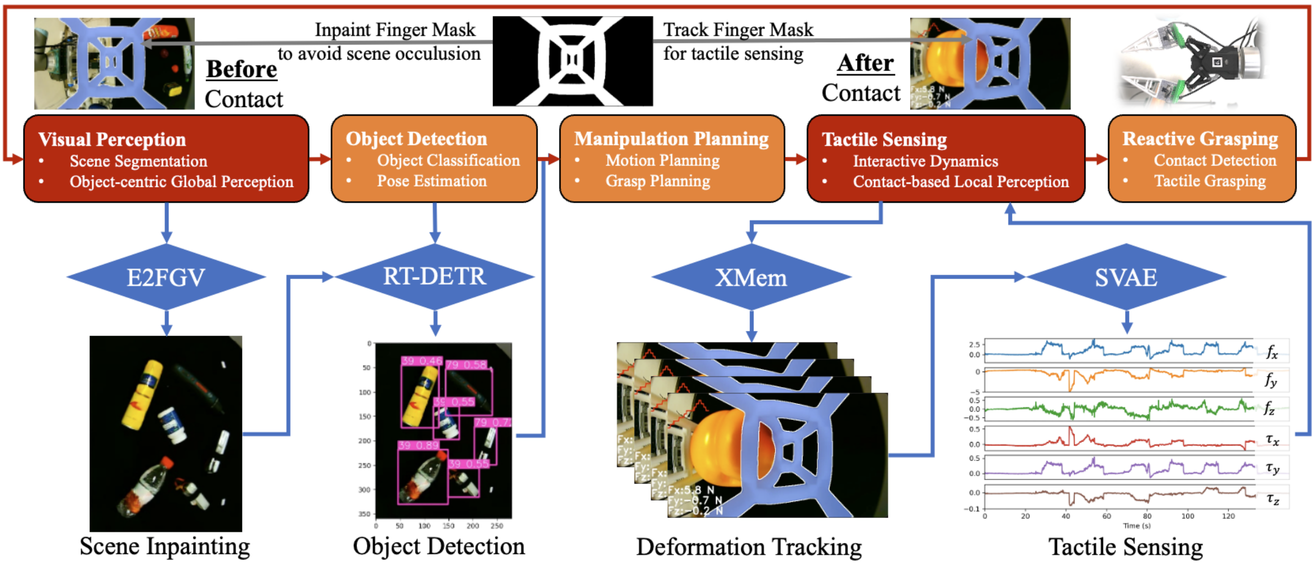

Figure 1. SeeThruFinger's in-finger vision for learning visual perception and tactile sensing.

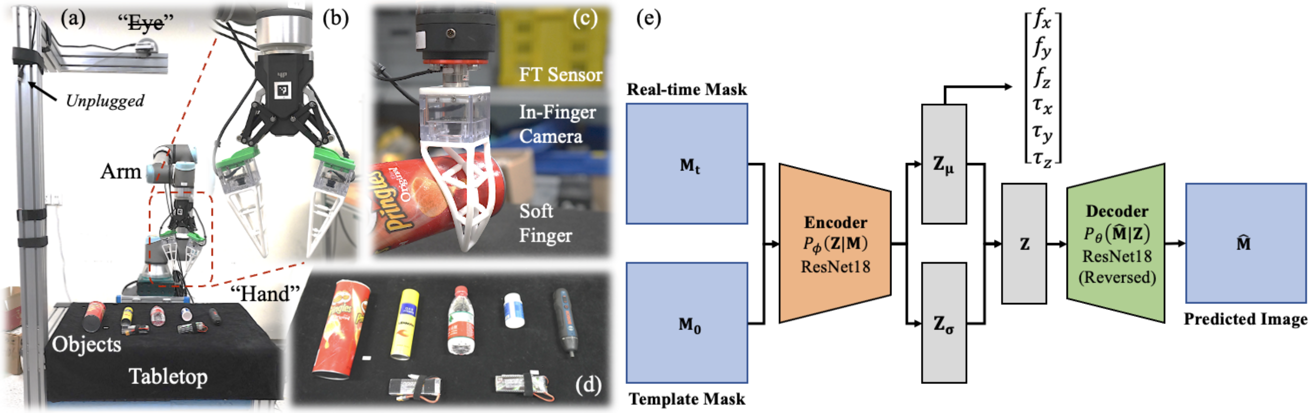

Figure 2. Experiment setup of the DeepClaw workstation and the network design for supervised variational autoencoder (SVAE). (a) The DeepClaw station used for this experiment uses the in-finger camera only with the external eye camera unplugged. (b) Detailed view of the gripper with the SeeThruFinger installed. (c) The SeeThruFinger setup for collecting training data against FT sensors, showing the finger adaptive to object geometries in spatial twisting, an in-finger camera installed inside a transparent housing, and a 3D-printed adaptor for customization. (d) The objects used for grasping later in this study. (e) The SVAE architecture for tactile learning.

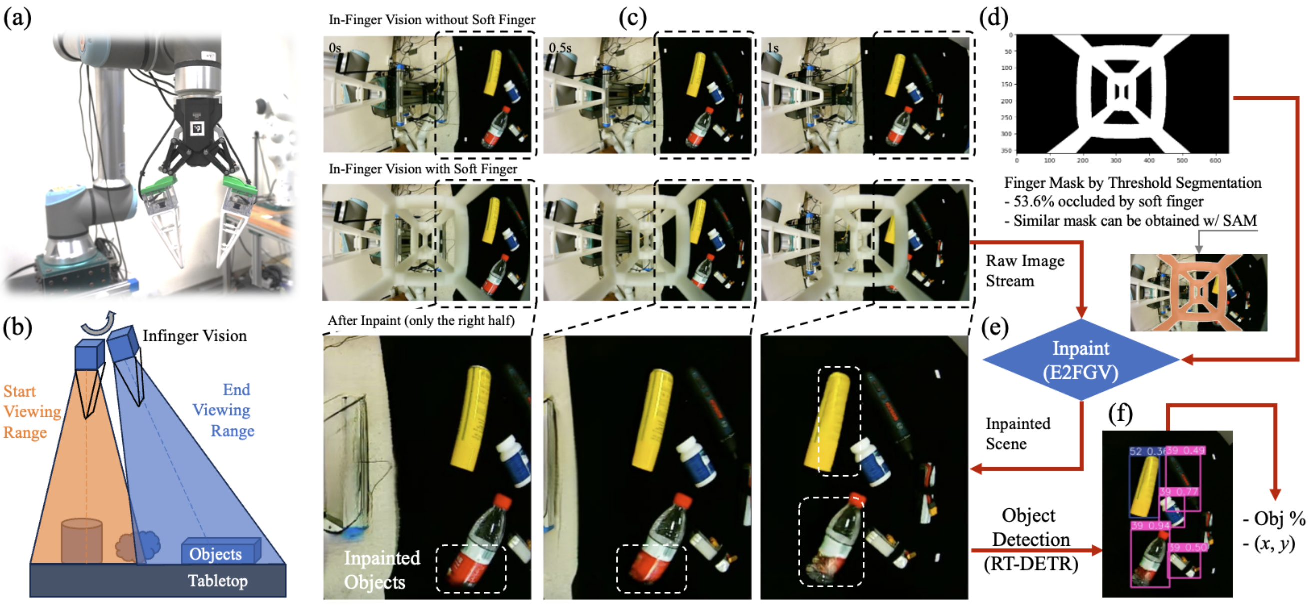

Figure 3. Scene inpainting using in-finger vision for object detection. (a) Full gripper assembly (Model AG-160-95 by DH-Robotics) with SeeThruFingers. (b) Illustration of scene scanning by slightly tilting the gripper above the tabletop to take a short video clip (1s) for inpainting. (c) Selected images at 0 s, 0.5 s, and 1 s from the head-up motion video clip without the soft finger, with the soft finger, and after inpainting in each row. (d) Finger mask obtained by threshold segmentation. (e) The inpainting operation. (f) Object detection using RT-DETR based on the inpainted scene.

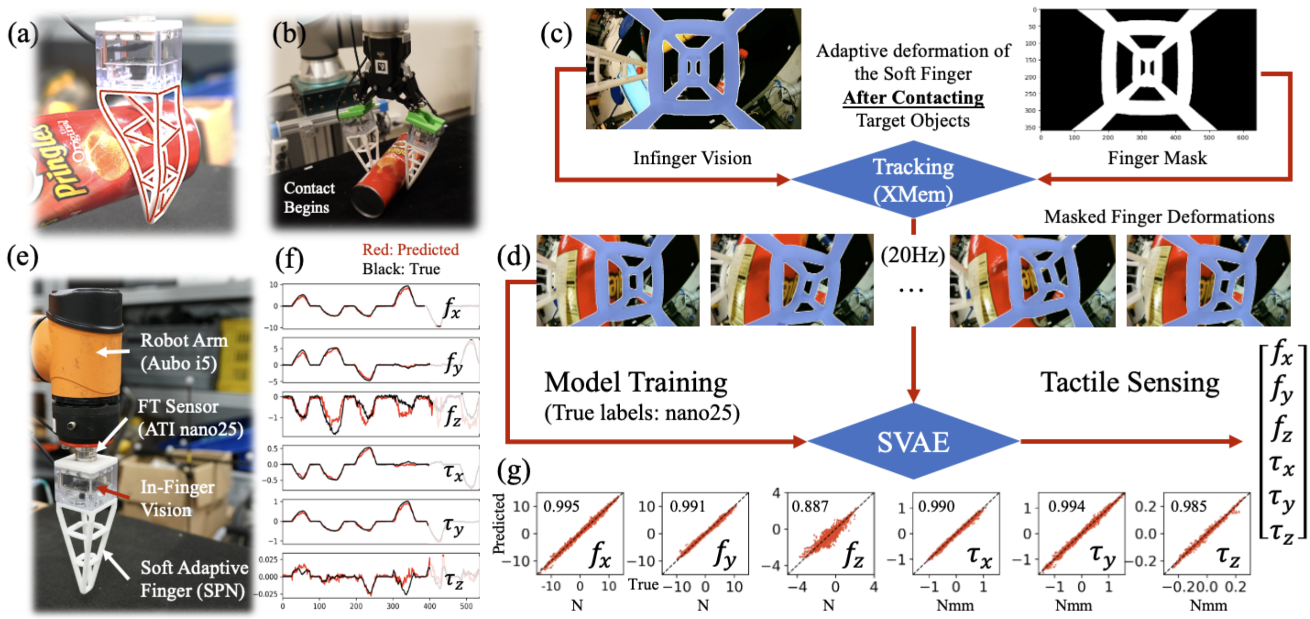

Figure 4. Tactile sensing using in-finger vision by tracking masks of the soft finger deformation after contact. (a) Omni-directional adaptation of the soft finger used in this study. (b) After contacting objects, the in-finger vision captures the soft finger's whole-body deformation. (c) Tracking with in-finger vision using XMem and the finger mask to generate segmented image streams of the soft finger's whole-body deformations for tactile sensing. (d) Tactile sensing using SVAE, with the left side showing the model training process and the right side showing hardware deployment, producing 6D forces and torques as outputs for reactive grasping. (e) Experiment setup for collecting training data for SVAE. (f) Testing results of SVAE for 6D forces and torques in sequence number. (g) Quantitative comparison between the predicted values and true labels obtained from nano25.

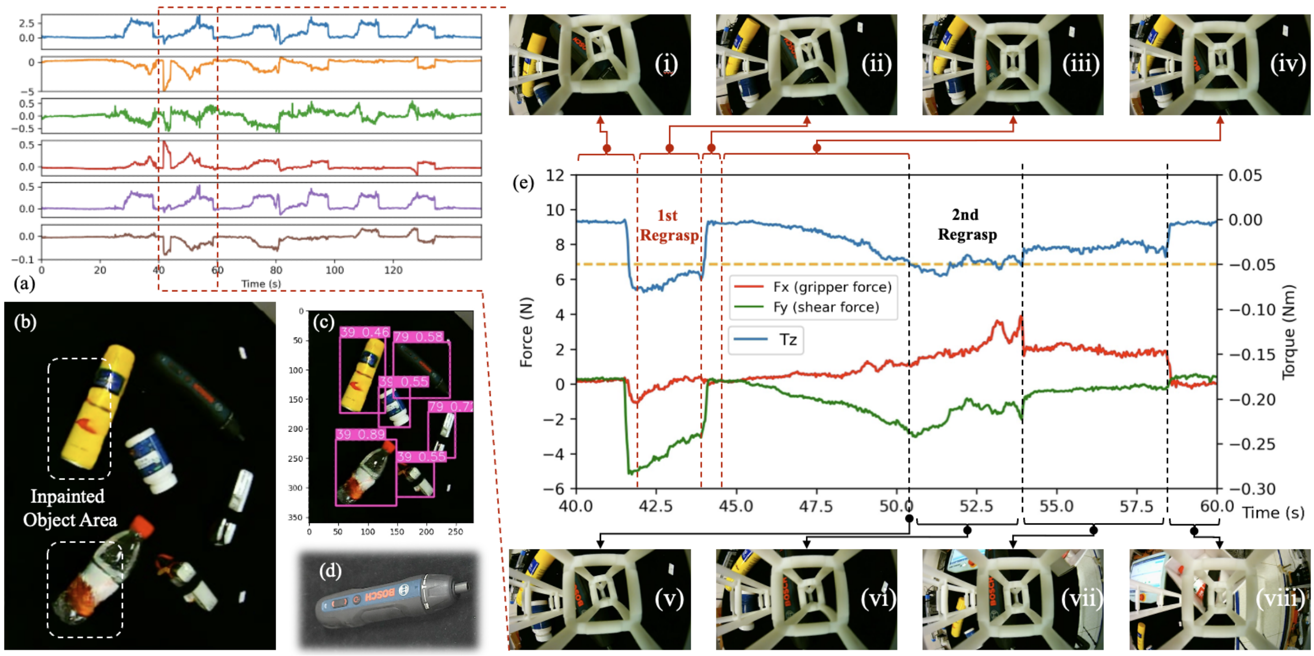

Figure 5. Learning visual perception and tactile sensing using SeeThruFinger for reactive grasping. (a) Results of 6D forces and torques recorded for reactive grasping. (b) Inpainted scene using E3FGV via in-finger vision, where the dashed boxes are inpainted portions of the objects. (c) Results of object detection using RT-DETR based on the inpainted scene. (d) The Bosche electrical screwdriver. (e) Detailed plot of the 6D forces and torques for the reactive grasping of Bosche electrical screwdriver using the SeeThruFinger and multiple screenshots of in-finger vision.

List of Tables

Table 1. Evaluation of the SVAE model for in-finger tactile sensing.

MAE	Force (N)			Torque (Nm)		
	f_x	f_y	f_z	τ_x	τ_y	τ_z
Validation dataset	0.241	0.212	0.14	0.024	0.027	0.005
Test dataset	0.328	0.283	0.24	0.027	0.036	0.005